G. FERRARI

COMPUTER-AIDED LEXICOGRAPHY AND SANSKRIT LANGUAGE

1. The use of mechanical techniques in lexicography is, now, widespread throughout the world and a large experience has also been acquired in the domain of historical languages.

1.1 In order to begin, it is perhaps worthwhile clarifying, to some extent, the terminology used in this field by recalling the most important phases of a standard lexicographic procedure [1]:

a) the text is prepared, according to chosen standards, and translated into machine-readable form. The choice and the application of these standards (pre-edition of the texts) is a particularly important step as it determines the following developments and results;

b) the text is printed out and corrected as many times as is necessary in order to eliminate all mistakes;

c) the, then corrected, text is processed by a series of programs in order to obtain certain results, among which the most classical are lexica and concordances. These are not different from those produced in a traditional way. The concordances, that is the list of forms occurring in a text, each of them associated with the list of the contexts in which these occur, are very similar to the card-archive used by the traditional lexicographer and have the same importance, being an obligatory step in computerized lexicography.

1.2 In spite of the similarity in results, computer-aided lexicography, compared with traditional methods, offers some practical advantages. The most obvious of these is the possibility of processing a large quantity of data in a short time, thus enabling the researcher to collect much more information in a much shorter period of time than would be possible without the assistance of the computer.

---

1. An almost complete review of methods and possibilities with discussions, may be found in *Linguistica Matematica e Calcolatori - Atti del Convegno e della Prima Scuola Estiva Internazionale, Pisa 16/VIII-6/IX 1970* (ed. Zampolli) sections « Lexicologie-Lexicographie » and « Statistique-Stilistique ».

Another, less evident but more relevant, advantage is the capability of the computer to perform every given series of operations, such as contextualization and word extraction, without introducing any clerical error, so that, once phase *b*) is executed, all the following operations require no other proof-corrections [2].

The high level of precision is guaranteed also in the sense that certain given criteria are uniformly applied throughout the whole procedure of text processing and operations on the text are performed without deviations. In this sense, every intervention or inspection on the text has the character of exhaustiveness as it will affect all involved cases [3], without any exceptions. This special characteristic has attracted attention to the expediency of choosing the most general and rigid standards; this is also in view of a recursive use of lexicographic materials. This implies that every individual research gives the opportunity of creating an instrument which may then be used for other researches [4], yet without compromising the original one.

2. The creation and improvement of such standards is only partly a technical matter as some general linguistic concepts are directly involved and a theoretical position may easily affect a practical solution. Thus, the definition of linguistic unit, of lexical unit, and of morphological paradigm is the turning point both in the creation of the input standards and in the operation of lemmatization [5]. In this way, and largely owing to the pressure of computer techniques [6], applied lexicology has acquired its own independence, developed its own methods and has given back to general linguistics some basic ideas.

The concept of dictionary itself has been somewhat modified. Traditional dictionaries of historical languages such as Latin, Greek or Sanskrit are generally conceived as instruments for the interpretation of texts. Thus, a word is normally defined by its translations, most of

2. This peculiarity acquires vital importance in contextualization, which consists in dividing the given text in as many contexts (phrases, verses or simply fixed sets of words) as it has words (occurrences). This operation, if done by hand, would mean the copying of the text at least three or four times, with the possibility of introducing a large number of clerical errors.

3. i.e. all the cases that formally correspond to the parameters given for the research as no substantial comparison is possible by machine.

4. The researcher, who works by hand, generally copies on cards only those passages of a text which are relevant to his own research. Working with the computer, complete concordances will be always produced and these become useful instruments for many other researches. By a second step, those particular passages required will be selected.

5. For a definition of the concept of lemmatization see below 4.2.

6. The development of lexicology as an independent science is certainly not due to the use of computers in linguistics; it is perhaps a case of parallelism in which, nevertheless, the demands of computer-aided lexicography accelerated the process and directed it to the identification of certain well-defined problems. It is in this sense that this section of the paper must be understood.

which are obtained with the aid of etymology, and its special uses are described only in so far as they occur in the grammar. Dictionaries and grammars of these historical languages were produced, more or less, with an identical purpose in mind, i.e. of saving the language and of demonstrating its correct use[7]. This characteristic of the first studies was inherited by the modern ones up until the first decades of this century.

But dictionaries have also proved to be the basic instrument in the scientific knowledge of the lexicon of a language, or of subsets of it. Therefore, the lexical index of a single work, or author or period, which was once limited to those sets in which special linguistic uses were expected *a priori*[8], became the most important way of gaining knowledge in syntactic and stylistic uses or linguistic structures of a given age, literary school or literary genre.

Linguistic statistics, for their part, assert that every given text is a sample of a linguistic universe, an infinite which can be known only by approximation and accumulation of samples[9]. So, the fundamental dictionary of a modern language is meant to show no more than the most representative model of that language[10].

But, in the domain of historical languages, the reality is somewhat different. In fact, although the theoretical concept does not change, there is no real infinite as enrichment is only possible when new manuscripts are found. The idea of *thesaurus* is a development of this concept, i.e. the complete archive of the texts of a language; an idea which becomes realizable only with the employment of mechanical techniques. It is self-evident that once this archive has been accomplished, the knowledge of the given language will overlap with the linguistic material; so, while it is impossible to say that we know the universe, it is true that we have the best approximation ever possible.

Once a historical language is known to this extent, a very general knowledge of the system is possible and every change and development may be differently appreciated if seen from within this system. In other words, after a *thesaurus* is accumulated, the history of a language and of each single phenomenon of it can be traced in a definitive way without

---

7. The Sanskrit dictionary of Böthlingk is, perhaps, slightly different as traditional Indian grammars seem to be mixed with new methods for the collection of materials (« *Unsere Wörterbuch besteht aus zwei nicht vollständig in einander aufzulösenden Elementen...* », *Sanskrit Wörterbuch, Vörwort I,I.*).

8. As, for instance, Milligan's *Vocabulary of the Greek Testament.*

9. The idea, borrowed from statistics, applies to linguistics, both as a mathematical model and as a general non-mathematical concept. See the works of G. Herdan, P. Guiraud and Ch. Muller.

10. In this domain, the correctness of the model has not only a quantitive aspect, but is also a problem of choice of the appropriate samples, see the introduction to all the frequency dictionaries of A. Juilland, in comparison with the *Lessico di frequenza della lingua italiana contemporanea* by N. Bortolini, C. Tagliavini A. Zampolli.

any risk of fragmentation. Thus, historical linguistics is not merely limited to the history of a few isolated phonetical changes but may really deserve its name.

3. These general remarks are also easily applied to the Sanskrit language. Moreover, the artificial character of Sanskrit makes the use of fixed schemata and formulas more relevant than in other languages; thus, the observation of even the most trivial regularities and irregularities in these formulas acquires a special importance. What we mean is that hardly any expressions or uses are arbitrary or casual; they are the product of choices which conceal religious, philosophical and historical facts. Thus, in tracing back the history of a formula, every slight variant is greatly relevant. In other words, for Sanskrit, more than for other languages, the accumulation of a large or complete archive of texts is an important realization. On the other hand, Sanskrit was used in some special fields such as logic or grammar, which are either ignored or insufficiently described by traditional dictionaries.

4. However, the mechanical treatment of Sanskrit poses some difficulties.

4.1 A first problem is in the choice between the *saṃdhi* or the *padapāṭha* as the standard input form of a text.

In a traditional *saṃdhi* text, in which words are physically joined together, automatic identification of the words is impossible [11]. Thus, v. Nooten, in his attempt to work out the *Mahābhārata* concordances, transcribed the text in *padapāṭha* [12]. Special linguistic reasons led Bernhard to choose a hybrid form of input, in which certain *saṃdhi* phenomena were separated and others were left as they were [13].

On one hand, it seems to be scientifically more correct to keep the text in its original form as the tradition is to quote examples, i.e. contexts, in *saṃdhi*. If one has to study *saṃdhi* and *in-pausā* phenomena then both forms must be recorded [14].

On the other hand, the identification of words in the text may take place only in a *padapāṭha* version or in hybrid transcriptions such as that adpoted by Bernhard, or by Aufrecht, in his edition of *Ṛgveda* [15],

---

11. This would be possible only with the aid of a morphological parser; see M. Kay, *Automatic morphological and syntactic analysis*, in A. Zampolli (ed.), *Texts, linguistic structures, and computers*, Amsterdam (in preparation).

12. See B. A. van Nooten, *A mechanical concordance for a Sanskrit work*, in JAOS, LXXXIV, 1964, pp. 56-58.

13. F. Bernhard, H. Reul, F. Schulte-Tigges, H. Sunkel, *Erstellung von Konkordanzen zu Sanskrit-Texten durch elektronische Rechenanlagen*, *Linguistics*, 22 (1966), pp. 5-23.

14. Statistics on phonemes of Sanskrit cannot be complete without a comparison between *saṃdhi* and *padapāṭha* figures.

15. In which, nevertheless, junctions of the type *nāsti* are not and cannot be solved without making a plain *padapāṭha* transcription.

in which words are physically separated, yet, still maintain their *saṃdhi* form. However, this choice does not satisfy the condition of keeping the original *saṃdhi*, while the use of *saṃdhi* or semi-*saṃdhi* units makes lemmatization particularly heavy as it should be preceded by a normalization of *saṃdhi* forms [16].

Therefore, the most appropriate solution would be to give as input a strictly *saṃdhi* text, provided with the information necessary in order to obtain a *padapāṭha*; this solution would have some additional advantages.

For some passages there is more than one traditional *padapāṭha* transcription; these may be assumed to be variants of a special tpye [17]. Now, although it is possible to record these using also a *padapāṭha* input, it is certainly easier to distinguish this type from normal variants if they refer to a single, basic text. Moreover, if new traditional readings are found or modern interpretations are given, the recording of them will be no more than a simple up-dating of a single archive. However, this alternative does not necessitate any additional work as, in all cases, one or more specialists are needed in order to prepare the text for input.

A system has been studied at CNUCE, Pisa, wich will provide the above-mentioned solution. It consists of providing each *saṃdhi* junction with an index by which the solution may be found in a matrix [18].

4.2 Another linguistically interesting phase of the lexicographic procedure is lemmatization. This consists in the reduction of all forms to a canonical one. This clearly implies a precise definition of that canonical form.

With reference to this subject, western and eastern lexicographic and grammatic schools have different traditions. In fact, the dictionaries of western languages tend to assume a form of the paradigm as lexical unit (lemma). So, for example, for Latin and Greek nouns, the singular nominative is assumed as the dictionary entry, while for Italian or French adjectives the masculine singular is the basic form.

---

16. If the three forms *devebhyaḥ*, *devebhyo* and *devebhyas°* are found in a text, they must be reduced to one canonical form before lemmatizing them to *deva-* (or *devaḥ* or *devas*, see below 4.2.). In fact, they are three realizations of a form that, under a morphological or statistical view-point is only one.

17. In fact, traditional *padapāṭha* transcriptions sometimes represent not only a plain interpretation of a passage given by a grammarian, but also the position of the philosophical school, to which that grammarian belongs.

18. This system was discovered by Dr. Michela Ott, who published, on this subject, a preliminary working paper, *Criteri preliminari di concordanza del vedico*, CNUCE, Pisa, 1976. We are now testing this technique in a project of automatic vedic concordances carried on by a cooperation between CUNCE-Divisione Linguistica and the Istituto di Glottologia of the University of Pisa, under the direction of Prof. R. Lazzeroni.

Sanskrit dictionaries, on the contrary, inheriting the long Indian tradition, assume as entries only stems or verbal roots; derived words are introduced in the course of the word description.

Some modern language dictionaries offer a third, intermediate alternative; that of recording the canonical form, and not a stem, only for primary words, showing derived ones as sub-entries.

Two choices, at two different levels, are identified here. The first one involves derivation; a morpho-semantic way of enriching the lexical wealth of a language. At this level, derived words may be considered either as morphologically produced and part of a single paradigm, or as new semantically autonomous units.

The second level involves flexion, i.e. the paradigm of each lexical unit, which may be represented either by a privileged form of the paradigm itself, or by a neutral element, i.e. the stem.

This latter alternative can be easily chosen without serious complications. On the contrary, choosing the former alternative at the first level, i.e. lemmatizing in the traditional Indian way would, sometimes, demand an interpretation or a comparison between traditional interpretations given by grammarians [19]. Moreover, a series of difficulties, such as how to treat compounds, arises. These considerations lead us to consider this choice as inopportune. In particular, this is because we are trying to furnish an instrument and we are not supposed to give preliminary interpretations to materials.

5. I have tried, here, to give a brief sketch of the advantages offered by the use of mechanical techniques in the field of Sanskrit lexicography. I have, also, tried to discuss some problems peculiar to the treatment of Sanskrit.

My purpose was to attract attention to the necessity of using the most rapid techniques in order to produce basic lexicographic works out of large quantities of linguistic material such as the Sanskrit corpus. To attain this, it is opportune to improve standard procedures so that they become suitable to the peculiarities of Sanskrit and its tradition.

It is self-evident, that no institution is so rich and so strong as to be able to carry on alone the project of a general « Sanskrit Thesaurus »; cooperation must be stimulated in order to achieve, as soon as possible, the following:

a) coordination of all lexicographic projects so as to avoid repetition;

b) a world standard procedure so as to allow the collection and exchange of materials, at any phase, without technical complications.

19. Interpretations, which otherwise could be comparatively assigned to each word assumed as a documentary instrument.